Molecular Therapy
Advances
Original Article

RAAVioli: A comprehensive approach to characterizing AAV vector integrations and rearrangements

Carlo Cipriani,[1,2] Laura Rudilosso,[2] Dhwanil A. Dalwadi,[3] Marco Masseroli,[1] Markus Grompe,[3] Eugenio Montini,[2] Andrea Calabria,[2,4] and Daniela Cesana[2,4]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy; [2]San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), IRCCS San Raffaele Scientific Institute, 20132 Milan, Italy; [3]Papé Family Pediatric Research Institute, Department of Pediatrics, Oregon Health and Science University, Portland, OR 97239, USA

In this study, we validated recombinant adeno-associated viral integration analysis (RAAVioli), a previously published computational method for the comprehensive analysis of adeno-associated virus vector (AAV) integration sites (ISs) and vector rearrangements across both long- and short-read sequencing platforms. Through *in silico* benchmarking and *in vivo* validation in a xenogeneic human hepatocyte model transduced with recombinant AAV (rAAV), we demonstrate the robustness and versatility of RAAVioli in accurately identifying AAV-genome junctions and reconstructing complex vector rearrangements across diverse experimental workflows. RAAVioli enables high-resolution detection of chimeric sequences, characterization of integration-site microhomologies and indels, and identification of complex vector structures, including multimeric concatemers. Short-read sequencing provided superior throughput compared to long-read sequencing, uncovering a larger number of integration sites and enabling precise quantification of clonal abundance, an essential parameter for evaluating the safety of gene therapy protocols. Benchmarking analyses further highlighted RAAVioli's higher sensitivity and accuracy relative to existing tools, establishing it as a reliable and scalable resource for AAV-based gene therapy and genome-editing studies. Moreover, RAAVioli's ability to profile structural rearrangements extends its utility to AAV vector characterization and quality assessment.

INTRODUCTION

Recombinant adeno-associated viral vectors (rAAVs) have been widely adopted as gene therapy delivery vehicles in preclinical and clinical studies.[1–9] Some AAV-based therapeutics have been also approved by both US Food and Drug Administration (FDA) and European Medicines Agency (EMA) for conditions like retinal disease, spinal muscular atrophy, and aromatic L-amino acid decarboxylase deficiency, while others, such as those for heart failure, myotubular myopathy, and glioma, are being fast-tracked for approval.[10] After transduction, rAAV primarily remains episomal within the nucleus of host cells but can undergo intra- and intermolecular recombination, resulting in monomeric circles or concatenated molecules with various structures like head-to-tail concatemers and more complex rearranged forms involving inverted or tandem repeats.[11–13] Although the rAAV genome and its rearranged forms can integrate into the host genome, this occurs at a low frequency (estimated at 0.1%–10% in hepatocytes).[14–18] The process relies entirely on host cellular machinery as rAAV vectors do not carry the proteins necessary for integration. Studies have demonstrated that double-strand breaks or nicks in the host genome promote AAV integration through homologous and non-homologous insertion mechanisms.[11,15–17] Despite the positive safety results obtained in clinical trials, the discovery that AAV can integrate into the host cell genome raises safety concerns, as a single intravenous systemic administration of AAV can transduce hundreds of millions of liver cells. Indeed, rAAV integration has led to hepatocellular carcinoma (HCC) formation in mice and is associated with clonal expansion of transduced hepatocytes in hemophiliac dogs.[12,15,19,20] A single case of HCC was also reported in a patient receiving liver-directed AAV5 therapy for hemophilia B,[21] and other reports have highlighted connections between human HCCs and natural AAV infections.[20,22] Therefore, given the potential risk of tumorigenesis associated with AAV integration, it is recommended to evaluate viral vector integration in the host genome in both animal models and patient samples during preclinical and clinical investigations, along with long-term follow-up studies.[10] Moreover, when rAAVs are used as DNA donor template in *in vitro* and *in vivo* gene editing applications, they integrate not only at the Cas-induced double-strand break but also at a myriad of other genomic locations where (Cas-independent) DNA damage has occurred. Such integration events have been observed in cultured murine cells, in human hematopoietic cells engrafted in immunodeficient mice, and following various *in vivo* gene editing protocols.[9,23,24] Hence

Received 20 June 2025; accepted 18 December 2025;
https://doi.org/10.1016/j.omta.2025.201659.

[4]These authors contributed equally

Correspondence: Daniela Cesana, SR-Tiget, IRCCS San Raffaele Hospital, Via Olgettina, 60, 20132 Milan, Italy.
E-mail: cesana.daniela@hsr.it

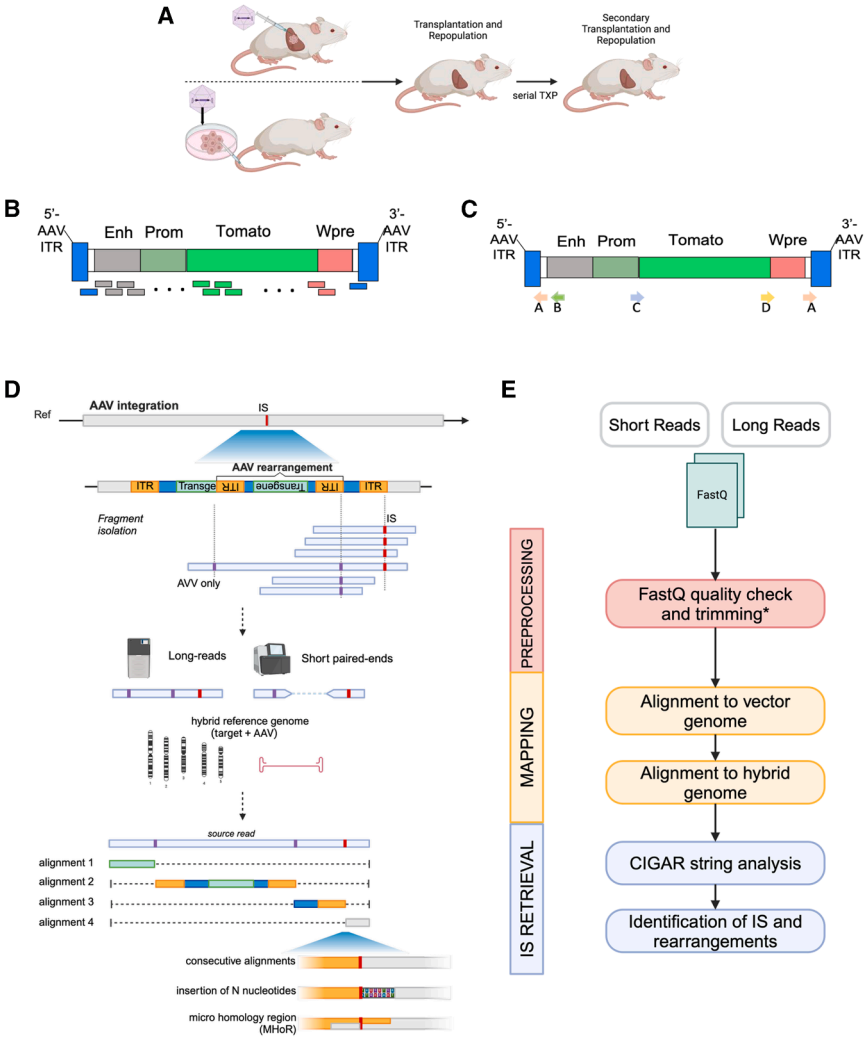

Figure 1. Retrieval of AAV IS

(A) Experimental strategy: human hepatocytes transduced *ex vivo* with AAV-CAG-tdTomato and then transplanted in FDH mice or transduced *in vivo* after liver regeneration in FDH mice; (B) position of the probes for AAV genome capturing: probes spanning the entire vector genome were used to select and capture genomic fragments containing the AAV portion, which were then utilized for long-read sequencing; (C) PCR primers designed at different positions of the AAV vector for SLiM-PCR: a total of four primer sets (A, B, C, and D) covering the entire vector genome were employed. Colored segments indicate the location of each individual primer set on the vector genome; (D) schematic flow for the identification of AAV integration and rearranged form; (E) RAAVioli pipeline workflow.

a thorough characterization of AAV integration sites (ISs) is essential even in these gene therapy contexts for safety assessment.

The identification of AAV ISs presents several challenges, due to the low frequency of integration events, the abundance of episomal vector forms, and the frequent occurrence of genomic rearrangements. Recent observations showed that AAV vectors may display complex rearrangement forms after integration,[14,25] highlighting the need for molecular and bioinformatic tools that can comprehensively identify and characterize all these events. Several tools have been developed for viral integration analysis, but they differ substantially in purpose, scope, and performance. Tools such as VIRUSBreakend,[26] ViFi,[27] and VERSE[28] are primarily designed to detect wild-type viral integrations in cancer or infectious disease contexts using whole-genome or large-scale high-throughput sequencing data. While powerful in those settings, these tools lack the sensitivity required to detect rare integration events typical of gene therapy studies and do not address vector-specific features such as rearrangements, episomal contamination, or low-abundance clones. Similarly, Isling[29] supports partial insertions and multiple viral types but does not enable structural reconstruction or analysis of complex rearrangements. Even within the narrower space of vector-specific pipelines, current tools remain limited. INSPIIRED,[30] ITR-Seq,[31] and VSeq-Toolkit[32] are optimized for integration site detection from PCR-based short-read data but do not resolve rearranged vector structures. INSERT-seq[33] enables analysis of long-read Nanopore data but lacks rearrangement characterization. Other notable studies, such as those by Farina et al.[7] and Nguyen et al.,[34] reported AAV integration and structural variants but relied on custom pipelines that do not offer a unified or reproducible analytical framework across different sequencing platforms. To address these limitations, we developed RAAVioli (recombinant adeno-associated viral integration analysis), a dedicated software suite designed to support both short- and long-read data and provide comprehensive profiling of AAV integration events. RAAVioli enables sensitive detection of chimeric reads, reconstruction of vector genome architecture (including concatemers and fragmented insertions), annotation of junctional features such as microhomologies and indels, and quantification of clonal abundance based on shear-site distribution, functionalities not offered by the aforementioned tools. Benchmarking analyses further demonstrate that RAAVioli outperforms one of the existing methods in both sensitivity and precision.

RESULTS

Methodological framework

To test and validate our experimental and analytical platform for the retrieval and characterization of AAV IS, we exploited a published[14] experimental design aimed at evaluating and characterizing AAV integrations in liver-directed gene therapy of human hepatocytes expanded in a mouse model of xenogeneic liver regeneration (Figure 1A).[14] In

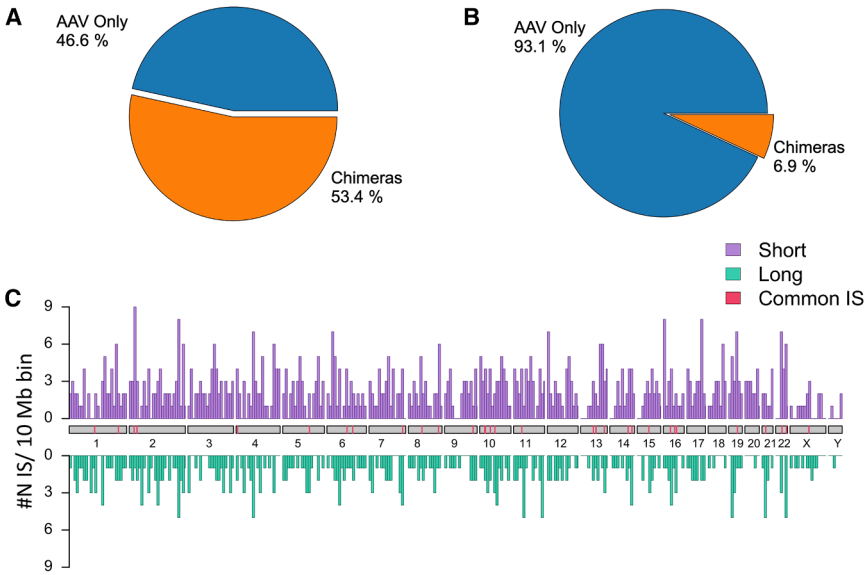

Figure 2. AAV chimerism in sequencing reads and genome-wide integration profile

(A and B) Pie charts showing the relative percentage of chimeric and AAV-only reads identified with the probe-based PacBio-long read approach (A) and with SLiM-PCR short-read method (B); (C) genome-wide distribution of the number of AAV ISs binned in 10 megabases of the human genome and identified with the short (upper panel) and long-read approach (lower panel) within the human genome. Human chromosomes are indicated; red bar within each chromosome indicates the position of shared IS, indicated as common IS.

this system, primary human hepatocytes were *in vivo* or *ex vivo* transduced with an rAAV expressing a tomato reporter transgene under the control of the CAG promoter and transplanted into immune-deficient mice lacking the fumarylacetoacetate hydrolase (Fah) enzyme to allow their expansion. In our previous study,[14] the rAAV ISs were identified by enriching the DNA isolated from transgene-positive hepatocytes for AAV sequences using transgene-specific probes and sequencing by long-range PacBio technology (Figure 1B). Here, the retrieval of AAV IS was performed starting from the same DNA material by adapting the Sonication and Linker-Mediated (SLiM) PCR. A sonication-based PCR method, conventionally used for clonal tracking and safety assessment in gene therapy clinical trials involving lentiviral vectors, gamma-retroviral vectors, and transposons, was adapted and combined with short-read sequencing approaches (Figure 1C).[35–37] Since AAV integrations within the host cell genome may originate from rearranged and fragmented forms of the vector genome, several PCR primers were designed to cover different portions of the AAV genome (including ITRs, promoter, and poly-A regions), allowing the retrieval of a substantial fraction of integration events involving both the vector ends and internal sequences. Overall, four different primer sets for PCR amplification were adopted for AAV IS retrieval, and all the obtained amplicons were sequenced with Illumina Novaseq platform (Figure 1C). We adapted the previous pipeline version used to identify and characterize long-reads to analyze the results of our PCR-based short-reads experiment setup. Overall, the computational approach for both long and short reads is based on aligning the reads to a hybrid reference genome, which is created by adding the vector sequence as an additional chromosome to the target genome (in this case, the human hg19 genome). This dual-reference strategy enables the identification and characterization of vector rearrangements and vector-host junctions by segmenting the reads into sub-alignments, which are then sorted and analyzed (Figure 1D). Specifically, sub-alignments from each read are parsed using custom scripts, allowing the reconstruction of events through the CIGAR string (Concise Idiosyncratic Gapped Alignment Report; Figure 1E). This method allows to identify chimeric reads containing fragments from both the vector and the reference genome, as well as reads that contain only rearranged or non-vector genomes, refereed as AAV-only sequences (Figures 1C and 1D; materials and methods). When RAAVioli was applied to both long- and short-reads, we found a significant difference in the number of chimeric reads between the two technologies (Figures 2A and 2B): 53.4%[14] for 3,280 AAV-containing reads from the long-read approach and 6.9% for 101,416,703 AAV-containing reads from the short-read approach intrinsically dependent on the shorter size of the read and the insert-size between the paired ends in which the genomic fragment can land. This may indicate a lower efficiency for IS identification by the short-read method, which may be compensated by the higher throughput and the lower cost per sequencing run.

Analysis of integration events in human hepatocytes

The position of each IS was defined within chimeric reads at the breakpoint between AAV and human genomic sequences (materials and methods). After collapsing chimeric reads with the same IS and structure, we identified 726 ISs from the short-read approach (Table S1). Depending on the PCR system adopted, the number of AAV ISs retrieved was variable, revealing that the highest number of ISs was achieved by the PCR system targeting the ITRs, while the lowest number was obtained from the system targeting the polyadenylation site (Figure S1A). Indeed, a total of 664 unique ISs were retrieved by combining the two ITR-targeting primer systems (351 ISs with primer A and 313 ISs with primer B, of which 24 were shared between the two), along with 68 and 18 ISs using primers C and D that target the promoter and the poly-A region of the expression cassette, respectively. Analyzing the DNA sequence at the junction breakpoint, we observed that approximately 43.4% of the IS exhibited a precise homology breakpoint (±1 nucleotides) between the vector integration point and the host genomic sequence, while the remaining were characterized by short stretches of nucleotides (up to 29 nucleotides) that could be mapped equally to both the provirus and the human genome or none (Figure S1B–S1E). These findings align with other studies indicating the presence of micro-homology regions (MHoRs) and nucleotide insertions (referred

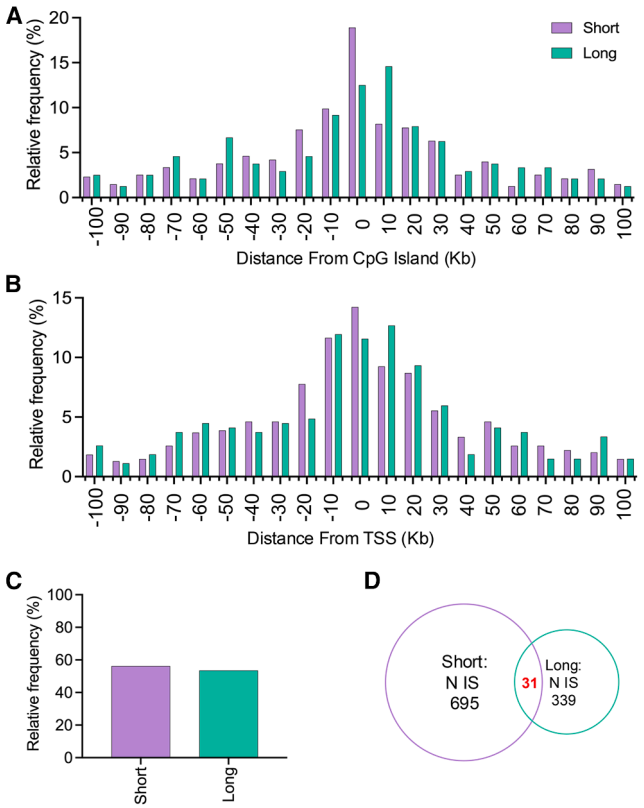

Figure 3. Genomic landscape of AAV integration sites

(A and B) Distribution of AAV IS identified with the long- and short-sequencing approach near CpG Islands (A) and TSS (B); (C) relative percentage (%, *y* axis) of AAV ISs targeting human RefSeq genes; (D) pie chart referring to the number of AAV ISs identified by the long- and short-sequencing datasets and indicating those that were commonly found by the two approaches.

to as "Insertions") or deletions at the integration loci.[11,14,16,17] Indeed, insertions and deletions were identified in 7% of the 370 ISs retrieved from these samples using the PacBio long-read approach.[14]

The genome-wide profile did not show any significant difference in the distribution of rAAV ISs across the human genome, regardless of the experimental method used, underscoring the reliability of both approaches (Figure 2C). In line with previous findings, these integration events showed a strong preference for integration near CpG islands, transcription start sites (TSSs), and RefSeq genes (Figures 3A–3C). Notably, 31 ISs were identified in both the long- and short-read IS datasets, further supporting the robustness of RAAVioli (Figure 3D). Analysis using the Genomic Regions Enrichment of Annotations Tool (GREAT)[38] revealed no biases toward specific gene classes in either dataset (data not shown). Of note, albumin, one of the most highly expressed genes in hepatocytes, was one of the most frequently targeted in both datasets (four ISs in the long-read dataset and three ISs in the short-read dataset; Table S2), further highlighting the consistency between the two approaches.

Clonality of the transduced populations of hepatocytes

Clonality analysis was conducted by examining the number of distinct fragment lengths associated with a specific IS. The overall assumption is based on the idea that two or more fragmented DNAs from sonicated independent molecules harboring the same integration are unlikely to share identical fragment sizes.[30,39] Therefore, the presence of multiple fragment lengths at the same IS suggests clonal expansion/proliferation of hepatocytes following the insertion event. We chose this method of quantification based on our previous experience with the retrieval and quantification of IS from integrative vectors.[35,40–42] We did not use unique molecular identifiers (UMIs) for clonal abundance quantification due to evidence of biases introduced by primer jumping during the amplification steps, which led to inflated clonal abundance estimates (Figure S2).

When comparing the abundance distribution of rAAV IS, a significantly higher number of genomes per IS was observed with the short-read approach compared to the long-read one ($p < 0.001$ by unpaired *t* test, average 12.7 ± 0.6 and 2.0 ± 0.04 genomes/IS; Figure 4A).

The distribution of IS abundances highlights the accuracy of short-read quantification methods compared to the PacBio long-read approach, as it correlates with the higher proliferation rates of human hepatocytes in the xenogeneic regeneration models. Human hepatocytes have a selective growth advantage in the *Fah*−/− environment, leading to the repopulation of the recipient's liver with human cells that undergo multiple rounds of division.[14] The division rate of transduced human hepatocytes is even higher than those observed from a previously published IS-derived dataset, which was obtained using the same PCR-based method and analyzed with the RAAVioli pipeline.[8] In the study by Calabria et al., adult *Mecp2*$^{-/-}$ mice were systemically injected with an rAAV vector. The liver showed an average of 4.3 ± 0.1 genomes per integration site ($p < 0.0001$ by unpaired *t* test; Figure 4A).[8] Notably, the 31 ISs shared between the short- and long-read datasets were derived from the largest clones of human hepatocytes as they were characterized by a significantly higher abundance compared to the entire IS population of the short-read dataset ($p = 0.007$ by unpaired *t* test, average 23.3 ± 6.4 genomes/IS; Figure 4). These data reflect the higher likelihood of retrieving these events from the entire population of transduced cells. To further characterize the shared IS, we analyzed their genomic annotations and compared their distribution against the non-shared IS, without significant readouts (data not shown). This result further corroborates that shared ISs are mainly determined by the relative abundance of clones rather than by biological preferences for specific genomic regions.

Rearrangements of the AAV vector

In addition to identifying AAV ISs, our platform enables the study and characterization of vector rearrangements in both AAV-only sequences and chimeric reads. Long-read sequencing methods revealed rearrangements of two or more AAV fragments in 74% of AAV-only sequences and 27% of integrating events (Figure 5A). In

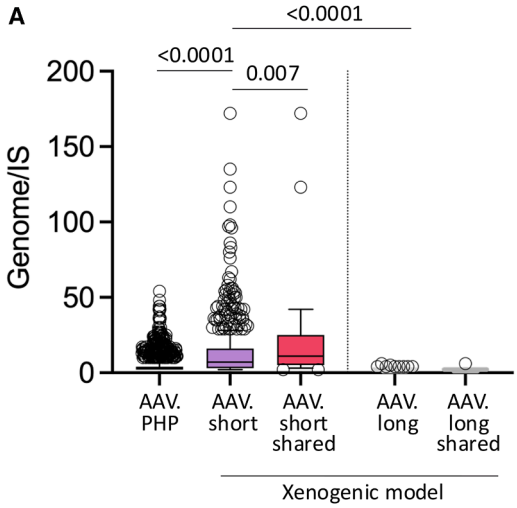

Figure 4. Abundance of AAV ISs

Cellular genomes for each AAV IS (genomes/IS ± standard error) identified among the different datasets as indicated: AAV-GT, IS identified in Calabria et al., Blood 2023, where adult Mecp2-deficient mice were systemically injected with an AAV.PHP vector; AAV-short, IS identified with short-read sequencing approach; AAV-long, IS identified long-read sequencing approach; Share-IS refers to ISs identified using both short- and long-read sequencing approaches. In the picture, "AAV-short shared" refers to the shared ISs quantified using the short read approach, "AAV-long shared" refers to shared ISs quantified with the long-read method. Statistics were performed by unpaired *t* test.

contrast, the short-read methods identified that only 33% of AAV-only and 8% of chimeric reads contained at least two AAV segments (Table S1; Figure 5A). While short-read sequencing is inherently limited in detecting multiple rearrangements, up to five vector fragments were observed in AAV-only sequences and two in chimeric reads (Figure 5A). Specifically, we observed that in the AAV-only dataset, different portions of the AAV genome can be involved in vector rearrangement, including those belonging to the plasmid sequence, such as those encoding for the Ori and the bacterial gene for resistance to the antibiotic ampicillin (Figure S3). On the contrary, in the AAV-chimeric dataset, most of the vector rearrangements involved the ITRs (58%) and the promoter region (27%). It is important to note that ITRs were analyzed collectively, without differentiating between the 3′ and 5′ ends, due to the potential alignment bias of short reads mapping exclusively within ITR regions, which can be ambiguously assigned between the two duplicated sequences.

Next, by examining the regions of the AAV genome most involved in integration events from the entire AAV-short read dataset, we observed that more than 80% of integrations involved the ITRs (Figures 5B and 5C), followed by the transgene (10%) and WPRE sequence (2%). A more detailed inspection can also reveal the preferred nucleotide breakpoints within the loop regions of ITRs (Figure 5D), as previously reported.[12,23,34] These results are in agreement with those obtained from the long-read datasets,[14] where ITRs were the most frequently identified at the junction breakpoints (63%), followed by the transgene coding sequence (15%), WPRE (13%), and promoter sequences (8%) (Figure 5E).

These results demonstrate that the frequency and types of rearrangements at the junction breakpoints were comparable between the two platforms,[14] thus supporting the use of short-read methods for the initial study of rearrangement events (Figure S3). Additionally, from a biological perspective, these data reinforce the tendency of rAAV to undergo complex backbone rearrangements and highlight the strong propensity of vector ITRs to be involved in integration events.[11,17,43,44] Consistently, short-read sequencing enabled near-complete coverage of the AAV vector genome, including plasmid backbone elements, reflecting the contribution of multiple rearranged AAV fragments captured across independent reads and consistent with the high recombination frequency typical of AAV genomes (Figure S3).

Benchmarking

To conduct an in-depth assessment of the pipeline, benchmark the precision, and recall for IS retrieval and characterization of vector rearrangements, we generated *in silico* reads mimicking real datasets, including both long and short reads, and we also modeled error rates at different thresholds (0.1%, 0.25%, 0.5%, 1%, and 2%) to model PCR artifacts (see materials and methods). We introduced a variable parameter k to define the acceptable distance between the expected and observed IS locations. RAAVioli achieved precision and recall values of 0.989 and 0.992 for IS identification and 0.945 and 0.998 for vector rearrangement characterization using the long-read approach and a k = 3 (Figure 6A). Similarly, when testing the performance of RAAVioli using the short-read approach, precision and recall values were 0.989 and 0.972 for IS identification (chimeras) and 0.950 and 0.997 for vector rearrangement characterization (AAV-only; Figure 6A). These results highlight the sensitivity and performance of RAAVioli in detecting insertion sites and vector rearrangements across different sequencing platforms. When evaluating RAAVioli's performance using simulated error models, our results showed an F1 score (a statistical metric that combines precision and recall results) greater than 0.9 across all tested scenarios (Figure S4; Table S3).

We then benchmarked RAAVioli's performance against ITR-seq,[31] a publicly available tool representing the current state-of-the-art for detecting AAV ISs using PCR-based methods and short-read sequencing approaches (Figure 6B). Different alignment quality filter parameters were used to adapt the test for the two pipelines (Table S3). RAAVioli outperforms ITR-seq in all the conditions tested, achieving an average F1 score of 0.98 compared to 0.93 when no quality filter parameters were applied.

Overall, the results highlight RAAVioli's ability to identify and characterize insertion sites and vector rearrangements using multiple sequencing platforms, outperforming existing tools when PCR-based and short-read sequencing approaches are adopted.

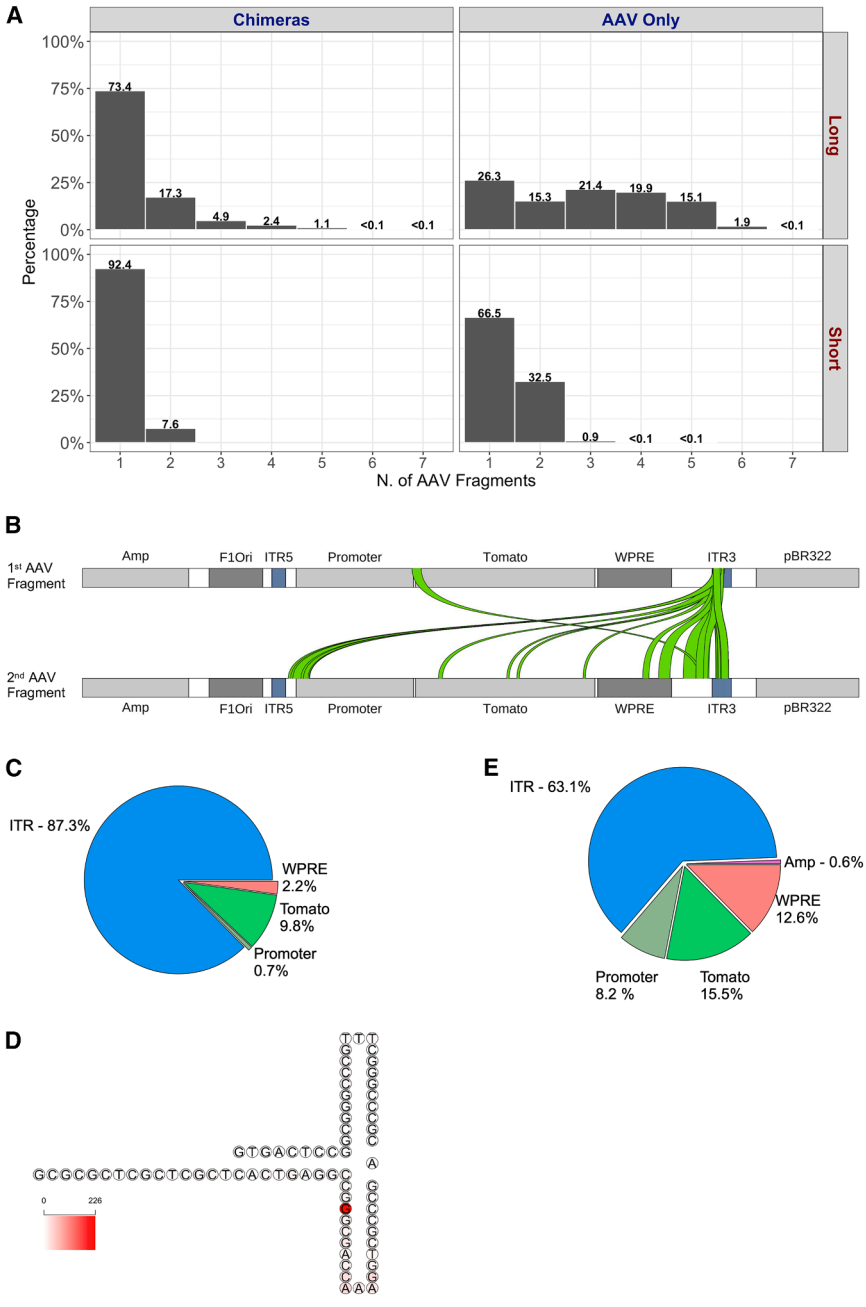

Figure 5. Rearrangements of the rAAV genome in human hepatocytes

(A) Percentage (*y* axis) of chimeric (left panel) and AAV-only (right) sequences identified with long- (upper panel) and short-read (lower panel) sequencing methods containing one to seven rearranged AAV fragments (*x* axis). (B) Structure of rAAV rearrangements in chimeric short reads. The first AAV fragment (top panel) is connected to the second AAV fragment (bottom panel) found in the chimeras. Most of vector rearrangement starts and ends in the ITRs; however, some rearrangement involved even internal portion of the vector genome. (C and E) Pie charts indicating the frequency of AAV features found at the IS position from the short- (C) and long-read IS dataset (E). (D) Heatmap of the AAV 3′ITR secondary structure, with the red scale indicating the frequency of AAV insertions occurring in the short-read dataset at the indicated nucleotide position.

DISCUSSION

In this study, we developed RAAVioli, a novel computational tool designed to comprehensively analyze AAV IS and vector rearrangements using long- and short-read sequencing approaches. RAAVioli provides a robust platform for detecting and quantifying chimeric sequences, characterizing integration locus microhomologies and indels, reconstructing sub-alignments to assess vector rearrangements, and identifying multimeric concatemers forms (head-to-tail, head-to-head, and tail-to-tail) from long reads.[14]

Our findings demonstrate that RAAVioli is a versatile and adaptable tool, capable of processing datasets from distinct experimental workflows and sequencing platforms. We validated RAAVioli by analyzing AAV ISs collected from the liver of a xenogeneic regeneration model of rAAV-transduced human hepatocytes,[14] using two distinct approaches: probe-selection methods with PacBio long-read sequencing[14] and sonication-based PCR approaches combined with short-read sequencing. Both methods can be employed for comprehensive and complementary analyses of AAV integrations, offering distinct advantages. Short-read sequencing generates a higher volume of reads, enabling the identification of a greater number of ISs compared to long-read sequencing. The observed 2-fold increase in IS recovery with the short-read approach emphasizes the utility of short-read sequencing in providing a higher throughput and broader coverage of integration events. RAAVioli's IS analysis confirmed several established genomic features of AAV integration, consistent across both datasets. AAV insertions were distributed sparsely throughout the human genome, following a pattern seen in previous studies.[14,16–18] A notable preference was observed for integration near CpG islands, TSS, and genes highly expressed in human liver, such as albumin.[14,16–18] This aligns with the current understanding that AAV vectors tend to integrate preferentially in transcriptionally active regions of the genome, especially in tissues where AAV is known to exhibit higher efficiency, such as the liver. The proximity of AAV ISs to CpG islands and transcription start sites (TSSs) suggests that integration may be influenced by epigenetic features, such as chromatin remodeling, that enhance

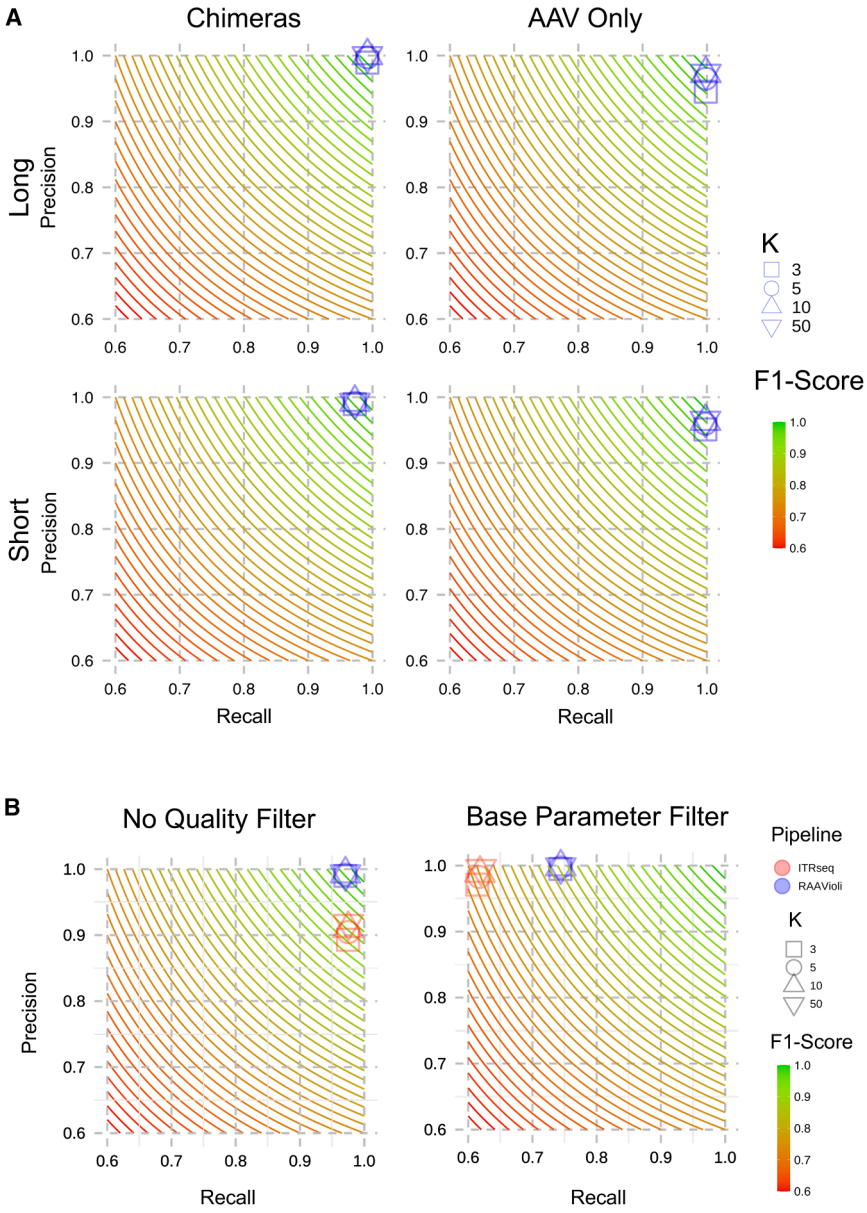

Figure 6. Precision and recall of RAAVioli and ITRseq

(A) Precision (*y* axis) and recall (*x* axis) scores of RAAVioli for the identification of ISs in simulated chimeric long reads (top left panel) and chimeric short reads (bottom left panel), as well as for the characterization of vector rearrangements in simulated AAV-only long reads (top right panel) and AAV-only short reads (bottom right panel). Different shapes represent different tolerance values (K) with mutation rate equal to 0. RAAVioli's performance remains consistent across tolerance thresholds. F1 score contours are shown as colored curves. Overall, RAAVioli achieves optimal performance, with F1 scores greater than or equal to 0.95 in all evaluated scenarios. (B) Precision (*y* axis) and recall (*x* axis) scores of RAAVioli (blue) and ITRseq (red) for the identification of ISs in simulated short reads, average score across different mutation rate. Different shapes represent different tolerance values (K). F1 score contours are shown as colored curves. Recall of ITRseq and RAAVioli drops when using mapping quality filter. Although for simulated reads those parameters are not relevant, they are important when applying the pipelines to real world data. Overall, RAAVioli shows higher performance than ITRseq in all conditions.

genome accessibility at these regions.[45] Interestingly, a substantial overlap of IS was observed between the two datasets. This overlap not only serves to cross-validate the two sequencing platforms but also demonstrates RAAVioli's versatility in handling data derived from distinct experimental workflows. The presence of these common ISs between the datasets further supports the robustness of RAAVioli as a tool for integration site discovery, highlighting its potential as a platform for comprehensive molecular analyses of AAV integrations.

In addition to providing a deeper evaluation of the overall clonal repertoire of transduced cells, the identification of ISs by sonication-based PCR approaches and short-read sequencing enabled precise quantification of IS clonality.[35] By leveraging the distribution of fragment lengths, this method outperforms UMI biases, offering a more accurate assessment of cellular abundance. Although UMI-based approaches are widely used in molecular genomics, we emphasize the importance of critically evaluating their performance within each protocol and assay to prevent hidden biases. Clonal abundance is a crucial feature for assessing the safety of gene therapy (GT) protocols in preclinical[34] and clinical studies,[35,41,42] as it may inform about the presence of potentially dangerous expansions due to gene deregulation caused by insertional mutagenesis, as shown in various clinical studies involving lentiviral and retroviral vectors.[35,42] Multiple cellular genomes were associated with the same AAV IS identified through the sonication-based approach. These numbers were significantly higher compared to AAV IS retrieved either from the same samples with the long-read approach or from the liver of adult mice that were systemically injected with a rAAV and analyzed using the same PCR-based approach and bioinformatics pipeline.[8] These results highlight the differences in proliferative capacity between transduced human hepatocytes in the xenogeneic mouse model and murine liver cells from adult mice following AAV injection[8,14] and reflect the sensitivity of our approach in quantifying the proliferation of transduced clones. Interestingly, AAV IS shared between long and short reads were also the most abundant in the short-read dataset.

Beyond IS characterization, RAAVioli effectively dissected AAV genome rearrangements, identifying integration patterns involving ITRs, transgenes, and regulatory sequences. Long-read sequencing demonstrated superior resolution for characterizing complex vector rearrangements and structural alterations, revealing up to six vector rearrangements in individual reads, while short-read data identified up to three. Long reads also enabled accurate reconstruction of large-scale structural variations, such as concatemers and folding, which short-read sequencing could not fully resolve. However, our results highlight that in both IS datasets, the ITR sequences of the AAV vector itself play a central role in the formation of IS in agreement with previous studies. This suggests that, rather than merely being passive elements in the vector genome, the ITR regions actively contribute to establishing IS, possibly through mechanisms related to the vector's structural integrity and the ability of ITR-containing fragments to recruit nuclear proteins involved in DNA repair that may facilitate integration events at single and double-strand breaks that are naturally occurring in the genome.[11] Additionally, short-read approaches generally exhibit lower error rates in base-calling compared to PacBio long-read sequencing, which results in more accurate base-pair-level information (Illumina 0.1%[46]; PacBio and Nanopore 0.4%[47]). This is particularly beneficial for detecting minor changes at the integration site, such as single nucleotide variants and small insertions/deletions, consistent with prior studies. Short-read sequencing captured preferential breakpoints within the loop regions of ITRs, enhancing our understanding of the structural integrity and recombination behavior of AAV genomes *in vivo*.

Benchmarking studies also highlighted RAAVioli's exceptional performance. *In silico* simulations confirmed its high sensitivity and precision for both IS retrieval and vector rearrangement characterization across long- and short-read datasets. Compared to existing tools, RAAVioli consistently outperformed in precision, recall, and F1-score metrics (a statistical measurement that combines precision and recall), underscoring its reliability and robustness.

The comprehensive integration analysis enabled by RAAVioli has broad implications for the use of AAV in gene therapy, supporting preclinical and clinical safety assessments while also providing insights that can improve our understanding of AAV biology in various gene therapy contexts. For instance, in a mouse model of AAV-intrathymic injection, we observed clusters of integrations within the RAG signal sequence of T cell receptor genes in circulating lymphocytes, a phenomenon resulting from the trapping of the AAV genome at RAG-mediated double-strand breaks produced during thymic T cell maturation.[8] Furthermore, RAAVioli was able to efficiently reveal unintended AAV insertions at nuclease on- and off-target sites in human hematopoietic stem cells *ex vivo* edited and engrafted in immune-deficient mice.[9] Additionally, RAAVioli's capacity to detect structural rearrangements makes this software a valuable tool for evaluating the quality of AAV-based drug products. Comprehensive pre-injection assessments of vector structural integrity could serve as a critical quality control measure, ensuring the safety of gene therapy treatments. Recent advancements in gene editing technologies, such as base editors and prime editors using AAV as the delivery vehicle, further emphasize the need for robust tools to characterize drug products and assess treatment safety comprehensively.[48] Although RAAVioli is primarily focused on AAV, its design makes it adaptable to other integrating elements, such as lentiviral or retroviral vectors, and transposons, showcasing its versatility across diverse applications. In the latter case, a representative application involves R2 transposons for which eukaryotic retroelement proteins have been repurposed to mediate targeted transgene insertion into human safe-harbor loci,[49] potentially resulting in various rearranged forms. RAAVioli could be used to characterize integration events with high specificity and efficiency, and with relatively modest sequencing depth, enabling analysis of bulk populations. Moreover, RAAVioli can potentially process long-read sequencing data from various platforms, including Oxford Nanopore Technology.[50]

In conclusion, RAAVioli bridges key gaps in the molecular characterization of AAV integrations by integrating multiple sequencing datasets with uniform criteria. This enables a holistic and high-resolution analysis of integration patterns, clonality, and structural variations, capabilities not achievable with previous tools. As such, RAAVioli represents a robust and versatile platform to support the development of safer and more effective gene therapy and genome editing strategies.

MATERIALS AND METHODS

Sample preparation

To assess AAV integration, we used a sonication-based linker-mediated PCR method (SLiM) commonly adopted to retrieve ISs from g-retroviral and lentiviral vectors. Briefly, genomic DNA purified from liver samples was sheared into ~1,000 bp fragments with a Covaris E220 Ultrasonicator, then end-repaired, adenylated, and ligated to custom linker cassettes containing an 8-nt barcode and 12 random nucleotides as Unique Molecular Identifier. The ligation products were amplified using 35 cycles of PCR with AAV-specific primers, followed by 10 cycles to add sequencing adapters, a second 8-nt barcode, and 12 random nucleotides. Finally, qPCR was used to quantify the amplified libraries prior to sequencing on Illumina NextSeq or NovaSeq platforms.

RAAVioli workflow for PCR-derived short-read sequencing

Preprocessing of sequencing reads

Raw sequencing reads were quality-checked using FastQC, and adapter sequences as well as PhiX reads were removed using Flexbar (v.3.5). Additionally, the first 12 random nucleotides were trimmed using Trimmomatic. Reads that did not begin with one of the predefined inner primer were discarded.

Alignment and mapping

The remaining reads were aligned to the AAV vector genome using BWA-MEM (v.0.7.17). Reads that successfully aligned to the vector genome were then mapped to a hybrid reference genome, consisting of the human reference genome (GRCh37) and the AAV vector genome. This dual-reference strategy enabled the identification of sequences spanning the vector-host junction, facilitating the detection

of rearrangements and integration events. To capture structural variations, all primary and supplementary alignments for each read were retained. RAAVioli natively supports all reference genomes, including human GRCh38 and any other species genome, provided in FASTA format with corresponding annotations.

Classification of reads and detection of rearrangements

The resulting BAM files were parsed using a custom Python script, which utilized the *pysam* library to collect all alignments for each individual read. For each read, alignments from the first read pair (R1) were sorted based on their CIGAR string, which describes the sequence of operations performed during alignment. This approach enabled precise characterization of each read. If multiple consecutive AAV alignments were present, the read was identified as containing a vector rearrangement, which could occur in both chimeric reads and AAV-only reads. Reads that contained at least one AAV alignment and one target genome alignment were labeled as chimeric reads. In this case, RAAVioli determined both the IS locus and the junction locus, defined as the last base of the vector adjacent to the IS locus. This step also enabled the detection of micro-homology regions (MHoR) and insertions (Ins) at the junction breakpoint. Reads in which R1 contained only AAV alignments were classified as AAV-only reads. The second end of the read (R2) is used for mapping purposes and to validate integration locus (the R2 must extend R1 target alignment that must be of at least 18 bp on the R1 and at least 30 bp when extended, e.g., an IS with an alignment on the target genome of 18 bp is considered valid only if the R2 extends the alignment to at least 30 bp; otherwise the read is discarded).

IS quantification and clustering

To quantify integration events, we employed a clustering strategy based on shear sites.[30,39] No lower limit of detection studies have been performed for the PCR-based platform, as the same technology has been already evaluated in another application of LV-based IS retrieval.[35] After this step, reads were grouped based on their IS locus, junction locus, and MHoR/Ins using hierarchical clustering (complete linkage, implemented in SciPy). Clusters were defined such that the maximum intra-cluster distance across these features was $\leq$20 bp (adjustable in RAAVioli). The abundance of each IS was estimated by summing the shear site ends across different triplets in the same cluster.

Post-processing and filtering

To minimize PCR artifacts and refine IS quantification, we applied the following filtering steps.

1. Low-abundance filtering: only integration events identified by at least two independent DNA fragments were considered in the analyses to remove potential PCR artifacts.
2. Collision filtering: ISs contained in more independent samples were removed using ISAnalytics (v.1.6.2).
3. Rearrangement-based filtering: ISs exhibiting the same rearrangement but different IS loci were removed, retaining only the most abundant occurrence.
4. Motif filtering: ISs associated with degenerated motifs were removed. This was performed using the MEME Suite, analyzing the genomic sequence 20 bp before the IS locus.
5. Final IS merging: ISs with loci separated by $\leq$ 10 bp were merged to correct for potential outlier reads from the initial clustering step.

Current limitation and future implementation steps

The current short-read branch of the RAAVioli pipeline is compatible only with PCR-based datasets, as it requires detection of vector-anchored primers for proper read orientation. In contrast, the long-read branch operates independently of primer detection and could therefore be adapted to process non-PCR-derived datasets, such as whole-genome sequencing (WGS) or targeted enrichment sequencing (TES). Furthermore, while the current version of RAAVioli uses a single default aligner for both long- and short-read data to ensure consistency and compatibility, its modular structure allows users to substitute alternative aligners, provided the output is in standard SAM/BAM format. Future versions may expand native support for additional alignment tools, particularly those optimized for long-read sequencing.

Finally, in its current version, RAAVioli supports a single vector reference per run, reflecting the design of our experimental datasets, which involved a single AAV construct. While multi-vector configurations are not yet implemented or validated, RAAVioli could potentially be adapted to support them by treating each vector as a distinct "chromosome" within the hybrid reference genome. Alternatively, integration site mapping in the presence of multiple AAV vectors can be performed by analyzing each vector separately or by concatenating vector sequences into a unified reference (removing duplicated ITRs if necessary), which may also enable detection of hybrid rearrangements. When PCR-based approaches are used, we recommend designing primers in vector-specific (ideally non-overlapping) regions to allow unambiguous assignment of sequencing reads to their corresponding vector.

Performance evaluation

To evaluate the performance of RAAVioli, we generated *in silico* datasets to assess its reliability in detecting both vector rearrangements and integration loci. A custom Python script was used to generate simulated short and long reads. These reads contained different numbers of AAV fragments as explained below.

- For short reads, the first AAV fragment consistently begins at one of several primer positions, replicating realistic scenarios:
  >ITR5 − chrV:1570–1590 −
  >ITR3 + chrV:5040–5060 +
  >OTHERV_1 + chrV:4660–4680 +
  >OTHERV_2 − chrV:3270–3290 −
  Both the first and rearranged AAV fragments follow a normal distribution with a mean of 70 and std 25. For the chimeric

dataset, we simulated a total of 200,000 reads containing varying numbers of AAV fragments: 175,000 reads with one fragment, 20,000 reads with two fragments, and 5,000 reads with three fragments. The fragment size distribution was modeled to reflect real experimental conditions. Each R1 read contains a minimum of 20 bp of genomic sequence. For the AAV-only dataset, we simulated a total of 45,998 reads containing a variable number of AAV fragments: 24,028 reads with two fragments; 16,701 reads with three fragments; 4,765 reads with four fragments; and 504 with five fragments. We next introduced error rates of 0.1%, 0.25%, 0.5%, 1%, and 2% to simulate realistic PCR artifacts, ranging from the nominal Illumina platform error rate (0.1%) to lower-quality reads (up to 2%), to evaluate pipeline performance under noisy conditions. Mutations were introduced into the reads using a custom Python script. This led to a total of 1,200,000 chimeric reads and 275,988 AAV-only reads being tested. Final plots were generated by averaging results across the different error rates (Figures 6B and S4).

- For long-read simulations, we generated 46,730 chimeric reads with an average length of 9,741.2 bp. Genomic fragment lengths were drawn from a normal distribution with a mean of 5,000 bp and a standard deviation of 1,500 bp. For each read, the AAV fragment lengths were sampled from a normal distribution, with the mean and standard deviation scaled according to the number of fragments simulated in that read. The distribution of AAV fragment counts per read in the chimeric dataset was as follows: 35,037 reads with one fragment; 8,426 with two fragments; and 3,267 with three fragments. For the AAV-only dataset, we simulated 29,986 reads. Fragment composition was as follows: 4,499 reads with two AAV fragments; 8,695 with three fragments; 8,397 with four fragments; and 8,395 with five fragments. We next introduced error rates of 0.5%, 1%, and 2% to simulate error rate to evaluate pipeline performance under noisy conditions. Mutations were introduced into the reads using a custom Python script. This led to a total of 186,920 chimeric reads and 119,944 AAV-only reads being tested.
- For chimeric reads, false negatives (FNs) were defined as reads erroneously classified as non-chimeric. True positives (TPs) were reads in which the insertion site was correctly identified within a tolerance of ±k bases from the simulated position to account for alignment shifts due to potential sequencing artifacts. False positives (FPs) were reads where the reported insertion site deviated by more than k bases from the simulated location. For AAV-only reads, TP were defined as reads in which all start and end positions of the simulated rearrangements were correctly identified within a tolerance of ±k bases. FPs were reads where the number and/or positions of the detected rearrangements did not match the simulated ones. FNs were reads that were incorrectly discarded.

Simulated reads were also used to compare RAAVioli performances to ITRseq,[31] which can identify AAV ISs in gene-therapy and gene-editing contexts. To ensure a fair comparison between RAAVioli and ITRseq, we ran both tools using their default parameters and by removing quality filtering thresholds. Regarding alignment strategies, a key difference between the two pipelines lies in the mapping tools used: ITRseq utilizes Bowtie2, while RAAVioli employs BWA-MEM. To account for this, we evaluated multiple combinations of quality score thresholds to robustly benchmark the performance of both pipelines (Table S3).

Statistical analysis

Statistical analyses were performed using GraphPad Prism 10, R software, and SciPy.

DATA AND CODE AVAILABILITY

Sequencing data have been uploaded in NCBI SRA (Sequence Read Archive) BioProject: PRJNA1347036. All relevant data are included in the manuscript. Integration site matrices derived from the PCR-based approach are provided in Table S1, while data from PacBio long-read sequencing are derived from Dalwadi et al.[14] The pipeline RAAVioli and its code is available at https://github.com/calabrialab/RAAVioli. The *in silico* datasets, their configuration files, and the configuration files to reproduce analysis on the generated short-read dataset are available at https://github.com/calabrialab/MTMRaavioli.

ACKNOWLEDGMENTS

This work was supported by Telethon Foundation (TGT11D1, TGT16B01, and TGT16B03). We are grateful to all members of the Montini lab for technical help and support, especially Fabrizio Benedicenti and Giulio Spinozzi. We thank Giulia D'Alessandro for the RAAVioli's logo.

The graphical abstract and Figure 1 were created in BioRender, https://biorender.com.

AUTHOR CONTRIBUTIONS

C.C. developed bioinformatics analyses, interpreted data, and wrote the manuscript. L.R. provided great technical support; D.A.D. performed the GT experiments on the xenogenic liver regeneration model and provided the DNA sample material; M.G. revised the manuscript and supervised gene therapy experiments on the xenogenic liver regeneration model; M.M. and E.M. revised the work and provide fruitful discussion; A.C. and D.C. conceived and supervised the project, interpreted data, supervised research, wrote the manuscript, and coordinated the work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used AI-assisted technologies to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omta.2025.201659.

REFERENCES

1. Wang, D., Tai, P.W.L., and Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nat. Rev. Drug Discov. *18*, 358–378. https://doi.org/10.1038/s41573-019-0012-9.

2. Colella, P., Ronzitti, G., and Mingozzi, F. (2018). Emerging Issues in AAV-Mediated In Vivo Gene Therapy. Mol. Ther. Methods Clin. Dev. *8*, 87–104. https://doi.org/10.1016/j.omtm.2017.11.007.

3. D'Avola, D., López-Franco, E., Sangro, B., Pañeda, A., Grossios, N., Gil-Farina, I., Benito, A., Twisk, J., Paz, M., Ruiz, J., et al. (2016). Phase I open label liver-directed

gene therapy clinical trial for acute intermittent porphyria. J. Hepatol. *65*, 776–783. https://doi.org/10.1016/j.jhep.2016.05.012.

4. Kaeppel, C., Beattie, S.G., Fronza, R., van Logtenstein, R., Salmon, F., Schmidt, S., Wolf, S., Nowrouzi, A., Glimm, H., von Kalle, C., et al. (2013). A largely random AAV integration profile after LPLD gene therapy. Nat. Med. *19*, 889–891. https://doi.org/10.1038/nm.3230.

5. Mattar, C.N.Z., Gil-Farina, I., Rosales, C., Johana, N., Tan, Y.Y.W., McIntosh, J., Kaeppel, C., Waddington, S.N., Biswas, A., Choolani, M., et al. (2017). In Utero Transfer of Adeno-Associated Viral Vectors Produces Long-Term Factor IX Levels in a Cynomolgus Macaque Model. Mol. Ther. *25*, 1843–1853. https://doi.org/10.1016/j.ymthe.2017.04.003.

6. Paneda, A., Lopez-Franco, E., Kaeppel, C., Unzu, C., Gil-Royo, A.G., D'Avola, D., Beattie, S.G., Olague, C., Ferrero, R., Sampedro, A., et al. (2013). Safety and liver transduction efficacy of rAAV5-cohPBGD in nonhuman primates: a potential therapy for acute intermittent porphyria. Hum. Gene Ther. *24*, 1007–1017. https://doi.org/10.1089/hum.2013.166.

7. Gil-Farina, I., Fronza, R., Kaeppel, C., Lopez-Franco, E., Ferreira, V., D'Avola, D., Benito, A., Prieto, J., Petry, H., Gonzalez-Aseguinolaza, G., and Schmidt, M. (2016). Recombinant AAV Integration Is Not Associated With Hepatic Genotoxicity in Nonhuman Primates and Patients. Mol. Ther. *24*, 1100–1105. https://doi.org/10.1038/mt.2016.52.

8. Calabria, A., Cipriani, C., Spinozzi, G., Rudilosso, L., Esposito, S., Benedicenti, F., Albertini, A., Pouzolles, M., Luoni, M., Giannelli, S., et al. (2023). Intrathymic AAV delivery results in therapeutic site-specific integration at TCR loci. Blood *141*, 2316–2329. https://doi.org/10.1182/blood.2022017378.

9. Ferrari, S., Jacob, A., Cesana, D., Laugel, M., Beretta, S., Varesi, A., Unali, G., Conti, A., Canarutto, D., Albano, L., et al. (2022). Choice of template delivery mitigates the genotoxic risk and adverse impact of editing in human hematopoietic stem cells. Cell Stem Cell *29*, 1428–1444.e9. https://doi.org/10.1016/j.stem.2022.09.001.

10. Sabatino, D.E., Bushman, F.D., Chandler, R.J., Crystal, R.G., Davidson, B.L., Dolmetsch, R., Eggan, K.C., Gao, G., Gil-Farina, I., Kay, M.A., et al. (2022). Evaluating the state of the science for adeno-associated virus integration: An integrated perspective. Mol. Ther. *30*, 2646–2663. https://doi.org/10.1016/j.ymthe.2022.06.004.

11. Deyle, D.R., and Russell, D.W. (2009). Adeno-associated virus vector integration. Curr. Opin. Mol. Ther. *11*, 442–447.

12. Miller, D.G., Petek, L.M., and Russell, D.W. (2004). Adeno-associated virus vectors integrate at chromosome breakage sites. Nat. Genet. *36*, 767–773. https://doi.org/10.1038/ng1380.

13. Nakai, H., Yant, S.R., Storm, T.A., Fuess, S., Meuse, L., and Kay, M.A. (2001). Extrachromosomal recombinant adeno-associated virus vector genomes are primarily responsible for stable liver transduction in vivo. J. Virol. *75*, 6969–6976. https://doi.org/10.1128/JVI.75.15.6969-6976.2001.

14. Dalwadi, D.A., Calabria, A., Tiyaboonchai, A., Posey, J., Naugler, W.E., Montini, E., and Grompe, M. (2021). AAV integration in human hepatocytes. Mol. Ther. *29*, 2898–2909. https://doi.org/10.1016/j.ymthe.2021.08.031.

15. Chandler, R.J., LaFave, M.C., Varshney, G.K., Trivedi, N.S., Carrillo-Carrasco, N., Senac, J.S., Wu, W., Hoffmann, V., Elkahloun, A.G., Burgess, S.M., and Venditti, C.P. (2015). Vector design influences hepatic genotoxicity after adeno-associated virus gene therapy. J. Clin. Investig. *125*, 870–880. https://doi.org/10.1172/JCI79213.

16. Nakai, H., Iwaki, Y., Kay, M.A., and Couto, L.B. (1999). Isolation of recombinant adeno-associated virus vector-cellular DNA junctions from mouse liver. J. Virol. *73*, 5438–5447.

17. Nakai, H., Montini, E., Fuess, S., Storm, T.A., Grompe, M., and Kay, M.A. (2003). AAV serotype 2 vectors preferentially integrate into active genes in mice. Nat. Genet. *34*, 297–302. https://doi.org/10.1038/ng1179.

18. Nowrouzi, A., Penaud-Budloo, M., Kaeppel, C., Appelt, U., Le Guiner, C., Moullier, P., von Kalle, C., Snyder, R.O., and Schmidt, M. (2012). Integration frequency and intermolecular recombination of rAAV vectors in non-human primate skeletal muscle and liver. Mol. Ther. *20*, 1177–1186. https://doi.org/10.1038/mt.2012.47.

19. Chandler, R.J., LaFave, M.C., Varshney, G.K., Burgess, S.M., and Venditti, C.P. (2016). Genotoxicity in Mice Following AAV Gene Delivery: A Safety Concern for Human Gene Therapy? Mol. Ther. *24*, 198–201. https://doi.org/10.1038/mt.2016.17.

20. Nault, J.C., Datta, S., Imbeaud, S., Franconi, A., Mallet, M., Couchy, G., Letouzé, E., Pilati, C., Verret, B., Blanc, J.F., et al. (2015). Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. Nat. Genet. *47*, 1187–1193. https://doi.org/10.1038/ng.3389.

21. Schmidt, M., Foster, G.R., Coppens, M., Thomsen, H., Dolmetsch, R., Heijink, L., Monahan, P.E., and Pipe, S.W. (2023). Molecular evaluation and vector integration analysis of HCC complicating AAV gene therapy for hemophilia B. Blood Adv. *7*, 4966–4969. https://doi.org/10.1182/bloodadvances.2023009876.

22. La Bella, T., Imbeaud, S., Peneau, C., Mami, I., Datta, S., Bayard, Q., Caruso, S., Hirsch, T.Z., Calderaro, J., Morcrette, G., et al. (2020). Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut *69*, 737–747. https://doi.org/10.1136/gutjnl-2019-318281.

23. Hanlon, K.S., Kleinstiver, B.P., Garcia, S.P., Zaborowski, M.P., Volak, A., Spirig, S.E., Muller, A., Sousa, A.A., Tsai, S.Q., Bengtsson, N.E., et al. (2019). High levels of AAV vector integration into CRISPR-induced DNA breaks. Nat. Commun. *10*, 4439. https://doi.org/10.1038/s41467-019-12449-2.

24. Nelson, C.E., Wu, Y., Gemberling, M.P., Oliver, M.L., Waller, M.A., Bohning, J.D., Robinson-Hamm, J.N., Bulaklak, K., Castellanos Rivera, R.M., Collier, J.H., et al. (2019). Long-term evaluation of AAV-CRISPR genome editing for Duchenne muscular dystrophy. Nat. Med. *25*, 427–432. https://doi.org/10.1038/s41591-019-0344-3.

25. Greig, J.A., Martins, K.M., Breton, C., Lamontagne, R.J., Zhu, Y., He, Z., White, J., Zhu, J.X., Chichester, J.A., Zheng, Q., et al. (2024). Integrated vector genomes may contribute to long-term expression in primate liver after AAV administration. Nat. Biotechnol. *42*, 1232–1242. https://doi.org/10.1038/s41587-023-01974-7.

26. Cameron, D.L., Jacobs, N., Roepman, P., Priestley, P., Cuppen, E., and Papenfuss, A.T. (2021). VIRUSBreakend: Viral Integration Recognition Using Single Breakends. Bioinformatics *37*, 3115–3119. https://doi.org/10.1093/bioinformatics/btab343.

27. Nguyen, N.P.D., Deshpande, V., Luebeck, J., Mischel, P.S., and Bafna, V. (2018). ViFi: accurate detection of viral integration and mRNA fusion reveals indiscriminate and unregulated transcription in proximal genomic regions in cervical cancer. Nucleic Acids Res. *46*, 3309–3325. https://doi.org/10.1093/nar/gky180.

28. Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. Genome Med. *7*, 2. https://doi.org/10.1186/s13073-015-0126-6.

29. Scott, S., Hallwirth, C.V., Hartkopf, F., Grigson, S., Jain, Y., Alexander, I.E., Bauer, D.C., and Wilson, L.O.W. (2022). Isling: A Tool for Detecting Integration of Wild-Type Viruses and Clinical Vectors. J. Mol. Biol. *434*, 167408. https://doi.org/10.1016/j.jmb.2021.167408.

30. Sherman, E., Nobles, C., Berry, C.C., Six, E., Wu, Y., Dryga, A., Malani, N., Male, F., Reddy, S., Bailey, A., et al. (2017). INSPIIRED: A Pipeline for Quantitative Analysis of Sites of New DNA Integration in Cellular Genomes. Mol. Ther. Methods Clin. Dev. *4*, 39–49. https://doi.org/10.1016/j.omtm.2016.11.002.

31. Breton, C., Clark, P.M., Wang, L., Greig, J.A., and Wilson, J.M. (2020). ITR-Seq, a next-generation sequencing assay, identifies genome-wide DNA editing sites in vivo following adeno-associated viral vector-mediated genome editing. BMC Genom. *21*, 239. https://doi.org/10.1186/s12864-020-6655-4.

32. Afzal, S., Fronza, R., and Schmidt, M. (2020). VSeq-Toolkit: Comprehensive Computational Analysis of Viral Vectors in Gene Therapy. Mol. Ther. Methods Clin. Dev. *17*, 752–757. https://doi.org/10.1016/j.omtm.2020.03.024.

33. Ivančić, D., Mir-Pedrol, J., Jaraba-Wallace, J., Rafel, N., Sanchez-Mejias, A., and Güell, M. (2022). INSERT-seq enables high-resolution mapping of genomically integrated DNA using Nanopore sequencing. Genome Biol. *23*, 227. https://doi.org/10.1186/s13059-022-02778-9.

34. Nguyen, G.N., Everett, J.K., Kafle, S., Roche, A.M., Raymond, H.E., Leiby, J., Wood, C., Assenmacher, C.A., Merricks, E.P., Long, C.T., et al. (2021). A long-term study of AAV gene therapy in dogs with hemophilia A identifies clonal expansions of transduced liver cells. Nat. Biotechnol. *39*, 47–55. https://doi.org/10.1038/s41587-020-0741-7.

35. Cesana, D., Calabria, A., Rudilosso, L., Gallina, P., Benedicenti, F., Spinozzi, G., Schiroli, G., Magnani, A., Acquati, S., Fumagalli, F., et al. (2021). Retrieval of vector integration sites from cell-free DNA. Nat. Med. *27*, 1458–1470. https://doi.org/10.1038/s41591-021-01389-4.

36. Jadlowsky, J.K., Hexner, E.O., Marshall, A., Grupp, S.A., Frey, N.V., Riley, J.L., Veloso, E., McConville, H., Rogal, W., Czuczman, C., et al. (2025). Long-term safety of lentiviral or gammaretroviral gene-modified T cell therapies. Nat. Med. *31*, 1134–1144. https://doi.org/10.1038/s41591-024-03478-6.

37. Magnani, C.F., Gaipa, G., Lussana, F., Belotti, D., Gritti, G., Napolitano, S., Matera, G., Cabiati, B., Buracchi, C., Borleri, G., et al. (2020). Sleeping Beauty-engineered CAR T cells achieve antileukemic activity without severe toxicities. J. Clin. Investig. *130*, 6021–6033. https://doi.org/10.1172/JCI138473.

38. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. https://doi.org/10.1038/nbt.1630.

39. Berry, C.C., Gillet, N.A., Melamed, A., Gormley, N., Bangham, C.R.M., and Bushman, F.D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics *28*, 755–762. https://doi.org/10.1093/bioinformatics/bts004.

40. Calabria, A., Spinozzi, G., Cesana, D., Buscaroli, E., Benedicenti, F., Pais, G., Gazzo, F., Scala, S., Lidonnici, M.R., Scaramuzza, S., et al. (2024). Long-term lineage commitment in haematopoietic stem cell gene therapy. Nature *636*, 162–171. https://doi.org/10.1038/s41586-024-08250-x.

41. Gentner, B., Tucci, F., Galimberti, S., Fumagalli, F., De Pellegrin, M., Silvani, P., Camesasca, C., Pontesilli, S., Darin, S., Ciotti, F., et al. (2021). Hematopoietic Stem- and Progenitor-Cell Gene Therapy for Hurler Syndrome. N. Engl. J. Med. *385*, 1929–1940. https://doi.org/10.1056/NEJMoa2106596.

42. Cesana, D., Cicalese, M.P., Calabria, A., Merli, P., Caruso, R., Volpin, M., Rudilosso, L., Migliavacca, M., Barzaghi, F., Fossati, C., et al. (2024). A case of T-cell acute lymphoblastic leukemia in retroviral gene therapy for ADA-SCID. Nat. Commun. *15*, 3662. https://doi.org/10.1038/s41467-024-47866-5.

43. Nakai, H., Wu, X., Fuess, S., Storm, T.A., Munroe, D., Montini, E., Burgess, S.M., Grompe, M., and Kay, M.A. (2005). Large-scale molecular characterization of adeno-associated virus vector integration in mouse liver. J. Virol. *79*, 3606–3614. https://doi.org/10.1128/JVI.79.6.3606-3614.2005.

44. Chang, H.H.Y., Pannunzio, N.R., Adachi, N., and Lieber, M.R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat. Rev. Mol. Cell Biol. *18*, 495–506. https://doi.org/10.1038/nrm.2017.48.

45. Hüser, D., Gogol-Döring, A., Chen, W., and Heilbronn, R. (2014). Adeno-Associated Virus Type 2 Wild-Type and Vector-Mediated Genomic Integration Profiles of Human Diploid Fibroblasts Analyzed by Third-Generation PacBio DNA Sequencing. J. Virol. *88*, 11253–11263. https://doi.org/10.1128/jvi.01356-14.

46. Fox, E.J., Reid-Bayliss, K.S., Emond, M.J., and Loeb, L.A. (2014). Accuracy of Next Generation Sequencing Platforms. Next Gener. Seq. Appl. *1*, 1000106. https://doi.org/10.4172/jngsa.1000106.

47. Dohm, J.C., Peters, P., Stralis-Pavese, N., and Himmelbauer, H. (2020). Benchmarking of long-read correction methods. NAR Genom. Bioinform. *2*, lqaa037. https://doi.org/10.1093/nargab/lqaa037.

48. Davis, J.R., Banskota, S., Levy, J.M., Newby, G.A., Wang, X., Anzalone, A.V., Nelson, A.T., Chen, P.J., Hennes, A.D., An, M., et al. (2024). Efficient prime editing in mouse brain, liver and heart with dual AAVs. Nat. Biotechnol. *42*, 253–264. https://doi.org/10.1038/s41587-023-01758-z.

49. Zhang, X., Van Treeck, B., Horton, C.A., McIntyre, J.J.R., Palm, S.M., Shumate, J.L., and Collins, K. (2025). Harnessing eukaryotic retroelement proteins for transgene insertion into human safe-harbor loci. Nat. Biotechnol. *43*, 42–51. https://doi.org/10.1038/s41587-024-02137-y.

50. Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K.F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. *39*, 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.